Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Permutation-based inference for spatially localized signals in longitudinal MRI data

Jun Young Park[a,*], Mark Fiecas[b], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Department of Statistical Sciences and Department of Psychology, University of Toronto, Toronto, ON M5S, Canada
[b] Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, MN 55455, U.S.A

ARTICLE INFO

Keywords:
Alzheimer's disease
Cortical atrophy
Permutation
Linear mixed effects
Spatially localized signals
Statistical analysis

ABSTRACT

Alzheimer's disease is a neurodegenerative disease in which the degree of cortical atrophy in specific structures of the brain serves as a useful imaging biomarker. Recent approaches using linear mixed effects (LME) models in longitudinal neuroimaging have been powerful and flexible in investigating the temporal trajectories of cortical thickness. However, massive-univariate analysis, a simplified approach that obtains a summary statistic (e.g., a p-value) for every vertex along the cortex, is insufficient to model cortical atrophy because it does not account for spatial similarities of the signals in neighboring locations. In this article, we develop a permutation-based inference procedure to detect spatial *clusters* of vertices showing statistically significant differences in the rates of cortical atrophy. The proposed method, called SpLoc, uses spatial information to combine the signals adaptively across neighboring vertices, yielding high statistical power while controlling family-wise error rate (FWER) accurately. When we reject the global null hypothesis, we use a cluster selection algorithm to detect the spatial clusters of significant vertices. We validate our method using simulation studies and apply it to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data to show its superior performance over existing methods. An R package for implementing SpLoc is publicly available.

1. Introduction

The spatial topography of the rate of cortical atrophy has been shown to differ between normal aging individuals and those diagnosed to have mild cognitive impairment (MCI) or Alzheimer's disease (AD), giving a quantitative characterization of regional changes in the cortex that accompany normal aging (Chan et al., 2003; McDonald et al., 2009). Structural magnetic resonance imaging (MRI) data collected longitudinally can be used to characterize the temporal trajectories of cortical thickness along the cortical surface defined on a manifold. This type of data gives rise to spatial data with measurements on vertices on a cortical mesh, as well as longitudinal trajectories of cortical thickness at each vertex.

The primary goal of this article is to develop a novel inferential procedure to detect spatial *clusters* of vertices where rates of cortical atrophy differ between those who are cognitively normal (CN) and subjects diagnosed with AD. While there have been a number of works in the neuroimaging literature that addressed similar research questions, in this article we focus on the methods based on linear mixed effects (LME) modeling because LME is powerful in detecting group differences using longitudinal neuroimaging data. Also, LME modeling is a flexible approach that can handle a non-uniform number of visits and missing rates, and it is useful in modeling longitudinal trajectories of a group, or group differences in trajectories, while accounting for the within-subject variability via random effects. In the functional MRI (fMRI) literature, Staffaroni et al. (2018) and Hart et al. (2018) used models based on the LME model to analyze longitudinal trends in functional connectivity (FC) networks in resting state fMRI. In the MRI literature, Bernal-Rusiel et al. (2013b) used an LME model to investigate the differences of the longitudinal trajectories of cortical thickness between subjects with AD and CNs. In imaging genetics, Xu et al. (2014) used an LME model for a longitudinal genome-wide association study (GWAS) for neuroimaging phenotypes. Fast and robust LME-based methods via the generalized estimating equation (GEE) have been developed in the neuroimaging literature (Guillaume et al., 2014; Liang and Zeger, 1986).

The vertex-wise LME (V-LME, Bernal-Rusiel et al. (2013a)) and spatiotemporal LME (ST-LME, Bernal-Rusiel et al. (2013b)) are two popular methods for analyzing longitudinal cortical thickness data and

[*] Corresponding author.
E-mail address: junjy.park@utoronto.ca (J.Y. Park).

[1] Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.neuroimage.2021.118312.
Received 12 October 2020; Received in revised form 11 May 2021; Accepted 23 June 2021
Available online 25 June 2021.

have been recognized in a number of applications (Gordon et al., 2018; Landin-Romero et al., 2017). V-LME fits a univariate LME to each vertex along the cortex. ST-LME is an extension of V-LME that fits a spatiotemporal model to a predefined subset of vertices and then adds a parametric spatial covariance structure to the model, which improves statistical power (Bernal-Rusiel et al., 2013b). Both V-LME and ST-LME conduct massive-univariate analyses to obtain vertex-wise p-values, and both then adjust for multiple comparisons by controlling for the false discovery rate (FDR) (Benjamini et al., 2006; Genovese et al., 2002). Both V-LME and ST-LME are readily available in FreeSurfer (Fischl, 2012).

Massive univariate analyses do not take advantage of spatial dependencies of the underlying *signal*, a term we use for the true value of a parameter of interest; in our case, the signal we are interested in is the difference in cortical atrophy between the CN and AD groups along the cortical surface. Furthermore, the vertex-wise effect size is low, justifying the need for smoothing the data during preprocessing (e.g., full-width at half-maximum [FWHM] of at least 8mm) to reduce the noise level (Coalson et al., 2018). However, excessive smoothing blurs the data, which in turn blurs the spatial extent of the underlying signal, resulting in a loss of spatial specificity. For example, a null vertex before any smoothing may become a pseudo-signal vertex after being smoothed when the null vertex is close to the signal vertex. Any statistical method that uses extensively smoothed data and identifies the pseudo-signal vertex as signal is prone to a loss of specificity. Thus, it would be necessary to choose a level of smoothing to the cortical thickness data that balances the decrease in the vertex-wise noise and spatial specificity (Bernal-Rusiel et al., 2010).

As discussed above, one major challenge in modeling longitudinal cortical thickness data is to improve power and sensitivity in detecting non-null vertices while simultaneously maintaining a reasonable range of smoothing that does not lead to a great loss in spatial specificity. On one hand, it is important to fully incorporate the spatial dependencies of the *signals* in the model as a form of cluster-wise inference in LME. For example, if there is a vertex showing a statistically significant difference in atrophy rate between those with AD and those who are CN, then it is also likely that neighboring vertices will also reveal statistical significance, in which the magnitude and the signs of the effect size is similar. On the other hand, it is beneficial to achieve high power by using rich spatial information and not depending on excessively smoothed data. This motivates a need for a new statistical method for cluster-wise inference without relying on Gaussian random field theory and its assumptions to justify the need for spatial smoothing (Lerch and Evans, 2005).

The main contributions of this article are to address the aforementioned challenges in modeling longitudinal neuroimaging data and provide a new permutation-based inference procedure with improved power and sensitivity relative to existing approaches. To this end, we expand on the burden test framework, which allows us to construct score test statistics that leverages the information from neighboring vertices to identify candidate spatial clusters in a data-adaptive way. Next, given the multiple comparisons problem that arises from the number of spatial clusters under consideration, we use a permutation procedure to control for the family-wise error rate (FWER), and the statistically significant clusters are picked using a cluster selection algorithm. The proposed method is computationally feasible since it requires fitting vertex-wise LMEs only and permutation does not require refitting LMEs. Also, the proposed method does not rely on excessive spatial smoothing during preprocessing, and we show this empirically by applying the proposed method in the analysis of longitudinal cortical thickness data with a minimal level of smoothing (FWHM= 2mm).

The rest of the article is organized as follows. In Section 2, we provide details to the proposed method. In Section 3, we evaluate the performance through simulation studies and apply SpLoc to the longitudinal cortical thickness data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We conclude with a discussion in Section 4.

2. Methods

2.1. Notation and model setup

Let $i = 1, \dots, n$ be the index for subjects, $j = 1, \dots, m_i$ be the index for visits for subject i, and $k = 1, \dots, s$ be the index for vertices defined by the triangulation of the cortical mesh. Also, let t_{ij} be the time between the baseline and the jth visit of the subject i. We let $\mathbf{x}_i$ be the q-dimensional covariate vector for subject i at the baseline and let $z_i = 1$ if subject i is in group 1 (CN) and -1 if subject i is in group 2 (AD).

We first present our model parametrization for each vertex based on the linear mixed effects (LME) models:

$$y_{ijk} = \alpha_{0k} + \mathbf{x}_i^T \boldsymbol{\alpha}_{1k} + z_i \beta_{0k} + t_{ij} \beta_{1k} + (z_i \cdot t_{ij})\gamma_k + b_{ik}^{(0)} + t_{ij} b_{ik}^{(1)} + \delta_{ijk}, \qquad (1)$$

where y is cortical thickness and the elements of $(b_{ik}^{(0)}, b_{ik}^{(1)})^T \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Phi}_k)$ are a random intercept and a random slope for subject i at vertex k. We let $\delta_{ijk} \sim \mathcal{N}(0, \tau_k^2)$ be the noise component and assume that the random effects and the noise are independent. The main parameter of interest in this paper is γ_k, a parameter for the interaction between time and clinical status, that allows us to compare the cortical decay rates between AD and CN. $\gamma_k = 0$ implies that two groups' baseline cortical thickness may vary (when $\beta_{0k} \neq 0$) but decay rates are the same. Therefore, setting $H_0(k) : \gamma_k = 0$ is a valid parametric null hypothesis for a vertex-wise difference between two groups' cortical decay rates. Extending this to the global null hypothesis $H_0 : \gamma_1 = \cdots = \gamma_s = 0$ enables brain-wise inference, with a two-sided alternative hypothesis H_1 : at least one $\gamma_k \neq 0$.

We now redefine the parameters in Eq. (1) for simplicity in notation. Let $\mathbf{X}_i$ be the matrix that contains of covariate information of an intercept, nuisance covariates ($\mathbf{x}_i^T$), and main effects of time and clinical status (z_i and t_{ij}) with the order of the visit j. Then, Eq. (1) is re-written as a simpler marginal form tailored for the null hypothesis:

$$\mathbf{y}_{ik} = \mathbf{X}_i \boldsymbol{\eta}_k + (z_i \cdot \mathbf{t}_i) \cdot \gamma_k + \epsilon_{ik}, \quad \epsilon_{ik} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{ik}), \qquad (2)$$

where $\mathbf{t}_i = (t_{i1}, \dots, t_{im_i})^T$. $\boldsymbol{\eta}_k$ in Eq. (2) is $(\alpha_{0k}, \boldsymbol{\alpha}_{1k}, \beta_{0k}, \beta_{1k})^T$ in Eq. (1) and ϵ_{ik} is $(b_{ik}^{(0)} + t_{i1} b_{ik}^{(1)} + \delta_{i1k}, \cdots, b_{ik}^{(0)} + t_{im_i} b_{ik}^{(1)} + \delta_{im_i k})^T$. In both equations, z_i and γ_k remain unchanged.

2.2. Statistical inference

2.2.1. Test for detecting spatially localized signals (SpLoc)

SpLoc provides an explicit cluster-wise inference framework that leverages spatial information to improve statistical power while controlling for FWER accurately and maintaining a reasonable computational cost. To achieve this goal, SpLoc first considers multiple spatial clusters (henceforth called *candidate clusters*) and computes a test statistic for each candidate cluster. As we will later see, SpLoc leverages power from multiple non-overlapping candidate clusters of signals where the effect size within each candidate cluster are similar.

We use vertex-wise test statistic from Eq. (2) as a building block for SpLoc. It is because Eq. (2) provides an explicit LME framework, where finding the maximum likelihood (ML) or restricted maximum likelihood (REML) is applicable to estimate parameters for every vertex k. In the LME framework, the Wald test, the likelihood ratio test (LRT), and the score test (also referred to as the Lagrange multipler test) are applicable to test $H_0(k)$, and they are all equivalent when the sample size is large (Buse, 1982; Cox and Hinkley, 1979). Among possible choices for vertex-wise test statistics, SpLoc uses the score test as a default for two major reasons. First, score-based testing in our setup is computationally efficient when using permutation because each permutation does not require refitting Eq. (2), improving computational efficiency. Please see Section 2.2.2 for details. Second, there have been a number of developments in statistics and neuroimaging recently that used the score test in adaptive association testing, which is useful in developing the SpLoc. Kim et al. (2014) and Ganjgahi et al. (2015) are other examples of relatively recent works that used the score test in neuroimaging studies.

To construct the score test for $H_0(k)$, consider parameter estimates $\{\tilde{\boldsymbol{\eta}}_k, \widetilde{\boldsymbol{\Sigma}}_{ik}\}$ from the model fitted under $H_0(k)$. The score test statistic for $H_0(k)$ from Eq. (2) is

$$U_k = \sum_{i=1}^{n} (z_i \cdot \mathbf{t}_i)^T \widetilde{\boldsymbol{\Sigma}}_{ik}^{-1} (\mathbf{y}_{ik} - \mathbf{X}_i \tilde{\boldsymbol{\eta}}_k), \tag{3}$$

which follows a normal distribution with mean 0 under $H_0(k)$. This score test statistic quantifies the expected change in model fit (quantified using the log likelihood constructed from Eq. (1) or (2); see Equation (5) of (Bernal-Rusiel et al., 2013b)) if γ_k were estimated using the data. Testing H_0 (i.e., no decay rate difference between AD and CN in any vertex) in the score test framework is followed by constructing the score vector $\mathbf{U} = (U_1, \ldots, U_s)^T$ and computing its covariance under H_0, denoted as $\mathbf{V}$, where $\mathbf{V}$ may be obtained in a closed form (with correct model specifications) or via permutation (see Section 2.2.2).

Various forms of test statistics using the score vector have been proposed in the generalized linear model (GLM) framework. Popular choices include the burden test (also referred to as the sum test), the variance component test, or adaptive methods that take the advantage of both the burden test and the variance component test (Kim et al., 2014). Among these, we focus on extending the burden test because it has attractive theoretical properties that motivate our primary purpose of detecting spatial clusters. The brain-wise burden test statistic adds all elements of the score vector and has the form of

$$T_{burden} = \frac{(\mathbf{1}_s^T \mathbf{U})^2}{\mathbf{1}_s^T \mathbf{V} \mathbf{1}_s} = \frac{(\sum_{k=1}^{s} U_k)^2}{\sum_{k=1}^{s} \sum_{k'=1}^{s} \mathbf{V}[k, k']},$$

where $\mathbf{1}_s$ is a vector of 1 with length s. When the sample size is large, the distribution of the test statistic is $\chi^2_{df=1}$ when the global null hypothesis is true. The burden test is motivated to provide high statistical power when all γ_k are the same across the cortex (Lee et al., 2014). This situation, however, is unlikely to happen in our motivating problem, and in fact we expect that the regions of different cortical decay rates are localized around specific brain regions.

Our proposed method extends the brain-wise burden test framework to construct a data-adaptive test. We first note that its extended version is well-motivated when signal locations are known *a priori*. Similar to the brain-wise burden test, we may consider adding score statistics in the signal locations only. Then, its power will be higher than (i) the brain-wise burden test that adds score statistic from signal and non-signal locations or (ii) the vertex-wise score test that does not consider all signal locations. One major limitation of this approach, however, is that the signal locations are unknown *a priori* and searching for the locations needs to be included at the expense of the loss of statistical power. To overcome the challenges, we construct an extended version of the burden test for possible candidate clusters, where each candidate cluster is defined by a vertex k and its neighboring vertices. Specifically, our proposed test statistic has a form of

$$T_{SpLoc} = \max\left\{ T_k^{(r)} = \frac{\left(\mathbf{w}_k^{(r)T}\mathbf{U}\right)^2}{\mathbf{w}_k^{(r)T}\mathbf{V}\mathbf{w}_k^{(r)}} \,\middle|\, k = 1, \cdots, s, \quad r \in \Omega \right\}, \tag{4}$$

where $\mathbf{w}_k^{(r)}$ is a binary vector (0/1) of length s and 1 is assigned only to the vertices that are within the r nearest neighbors of the vertex k (including the kth vertex itself). Therefore, $T_k^{(r)}$ is the burden test statistic for a specific cluster constructed by vertex k and its $r-1$ neighbors. Ω is the set of the sizes of neighbors that is specified by the user and provides a degree of flexibility. T_{SpLoc} takes the maximum of burden test statistics of all candidate clusters with varying sizes determined by Ω. Note that, when the sample size is large, the distribution of $T_k^{(r)}$ is $\chi^2_{df=1}$ under the global null hypothesis, regardless of the choice of vertex k and neighbor size r. In addition, it provides high power when the γ_ks are the same within the candidate clusters. Specifically, our approach is well-motivated when there are multiple non-overlapping spatial clusters of signals where the effect sizes (γ_k) within each spatial cluster are the same. Since both the number of the clusters as well as the true locations of the clusters are unknown, we consider various choices of neighbors for every vertex to scan for signal clusters. Because we have approximately 10k vertices in each hemisphere using the fsaverage5 template, we use $\Omega = \{1, 5, 10, 20, 30 \ldots, 100, 150, 200, 250, \ldots, 500, 600, \ldots, 1000\}$, though we point out that other choices for this set are possible. This choice of Ω incorporates a wide range of cluster sizes in the data analysis. Furthermore, this Ω has more candidates with small sizes to reflect the possibility of the existence of small signal clusters and to balance between sensitivity and specificity in the selection step (see Section 2.2.3). To illustrate the effects of different choices of Ω, consider the case when $\Omega = \{1\}$. In this setting, there is no neighboring vertex without each vertex itself and T_{SpLoc} reduces to the maximum of the vertex-wise score test statistics (i.e., the $\min P$ approach). Similarly, the case when $\Omega = \{s\}$ yields the test statistic for the brain-wise burden test (T_{burden}).

Our proposed method is data-adaptive because we combine test statistics of various candidate clusters through the $\max T$ approach. Because the test statistic for every cluster has the same distribution under the global null hypothesis, test statistics for larger clusters do not necessarily contribute to the power of the SpLoc. Similar to other cluster-wise inference methods in neuroimaging, our method is well-motivated to detect large signal clusters with low effect size or small signal clusters with high effect size.

2.2.2. *Permutation*

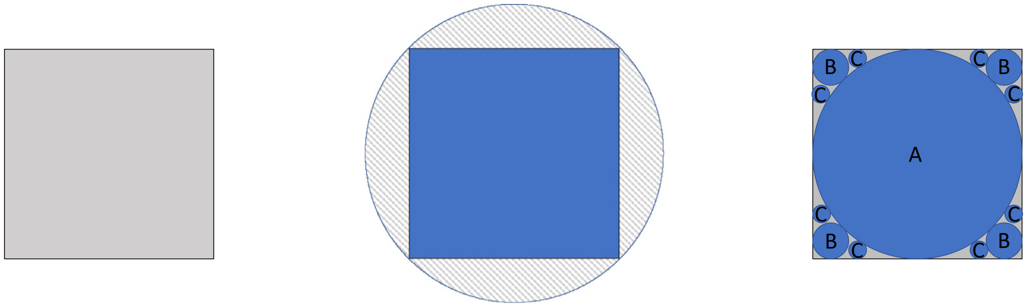
This section addresses how to compute the p-value of the SpLoc test via permutation. Permutation is especially helpful in our setting in three ways. First, while each burden test statistic follows $\chi^2_{df=1}$, the proposed test statistic T_{SpLoc} does not follow an explicit probability distribution under H_0. This is a typical situation in an adaptive testing procedures, and Monte Carlo and/or permutation is useful to control FWER in the weak sense (Kim et al., 2014). We note that control of FWER in the weak sense means that the global null hypothesis is assumed to be true, whereas control of FWER in the strong sense means that we may truly have any combination of nulls and non-nulls. Second, each element of the score vector $\mathbf{U}$ is computed vertex-wise without an explicit spatial covariance model for the cortical thickness data. However, cortical thickness data are spatially correlated and an estimate of $\mathbf{V}$ with the model assumption in Eq. (1) will lead to overly conservative results. Permutation provides a useful tool to compute $\mathbf{V}$ when the score vector is computed from the vertex-wise model.

In this section, we provide a permutation-based approach to obtain a better estimate of $\mathbf{V}$ and eventually to compute p-value. Note that each $H_0(k)$ tests on a specific parameter in linear model when there are nuisance variables added to adjust for confounders. There are a number of ways to conduct permutation testing, as summarized in Winkler et al. (2014). Our approach is analogous to Draper and Stoneman (1966), and, in terms of comparing group differences in LME, Braun and Feng (2001) to compute p-value of the SpLoc test. Due to how the score test statistic is constructed, the proposed method does not require refitting the model for each permutation, reducing computational cost dramatically. Our permutation algorithm is as follows:

1. Permute group assignments (z_i) and apply Eq. (3) to generate permuted score vector $\mathbf{U}^{(b)}$, $b = 1, \ldots, B$, where B is the number of permutations.
2. Compute $\mathbf{V}$ using the sample covariance of $\mathbf{U}^{(b)}$, $b = 1, \ldots, B$.
3. Apply Eq. (4) to compute T_{SpLoc} using $\mathbf{V}$ and $\mathbf{U}$ as well as $T^{(b)}_{SpLoc}$ using $\mathbf{V}$ and $\mathbf{U}^{(b)}$.
4. The p-value of SpLoc is the proportion that the permuted test statistic $T^{(b)}_{SpLoc}$ exceeds the original test statistic T_{SpLoc}.

2.2.3. *Cluster selection*

Our proposed test statistic is designed to test the global null hypothesis H_0, i.e., that none of the vertices of the brain shows statistically significant differences in cortical decay rates between those with AD



(i) True shape of the signal. (ii) Selection with a large candidate. (iii) Selections with small candidates.

Fig. 1. (i): The true shape and size of a signal cluster. (ii): By using a neighbor set that is larger than the signal, the selected signal cluster will have high sensitivity (blue square) but with low specificity (hatches). (iii): By including smaller choices to Ω, the selected signal clusters (the collection of non-overlapping A, B, C circles) will have lower sensitivity but higher specificity, provided that the test statistics for A, B, and C are above the threshold.

and those who are CN. Rejecting the global null hypothesis itself does not provide any further information on the signal clusters, such as how many, their locations, their sizes, and their shapes. Fortunately, the way we constructed the test statistic is related to a selection algorithm for detecting signal clusters. The selection algorithm is based on using the extended burden test statistic to determine the spatial clusters (defined by a vertex and its $r-1$ nearest vertices). The set of signal vertices can be identified following Jeng et al. (2010):

1. Set FWER (e.g., $\alpha = 0.05$). Set a threshold t^* to be the $100 \times (1-\alpha)\%$ quantile of the null distribution $\left\{T^{(b)}_{SpLoc} | b = 1, \dots, B\right\}$ that controls FWER.
2. If T_{SpLoc} is less than or equal to the threshold t^*, conclude that no vertex is statistically significant. If not, consider candidate clusters with test statistic greater than the threshold.
3. Choose the candidate cluster with the highest test statistic and identify all vertices within the cluster as signals. Remove other candidate clusters that overlap with the identified signals.
4. Repeat step 3 iteratively until there is no other candidate to consider.

We provide an illustration of the selection algorithm in Appendix A. The key idea is that each of the extended burden test statistic follows $\chi^2_{df=1}$ asymptotically under H_0, and we use the ordering of the test statistics to prune off clusters that consist mostly of noise vertices. As a special case, all clusters constructed by $\Omega = \{1\}$ only include every vertex only and SpLoc with $\Omega = \{1\}$ is equivalent to detecting vertices whose test statistic exceeds the threshold that controls FWER. The first two parts of the selection algorithm ensure that the signal clusters are chosen among the set of the candidate clusters whose test statistics exceed the threshold that controls FWER. The rest of the algorithm narrows down the threshold-passing candidate clusters.

The proposed algorithm has theoretical guarantees of the detected signal clusters under several assumptions, including that (i) the score test statistics are independent and (ii) each true signal cluster is one of the candidate clusters (Jeng et al., 2010). We point out that the first assumption does not hold with cortical thickness data since the data exhibit spatial correlations, but, as we will see later, we evaluate the proposed algorithm's performance on both simulated and empirical data. Furthermore, even though threshold by Jeng et al. (2010) is determined by random matrix theory under the independence assumption, we use a permutation approach to control for false positives, which would be valid in cortical thickness data with spatial correlations. Second, given s vertices one would be able to create $2^s - 1$ possible clusters, and thus using all possibilities as candidate clusters would be computationally impossible. Furthermore, the neighbor specifications in practice only provide an approximation to the true underlying signal. We show an example in Fig. 1 on how the choice for Ω can impact sensitivity and specificity.

Table 1
Comparison of the proposed method (SpLoc) to other LME-based methods in longitudinal neuroimaging data.

Methods	Spatial correlation of cortical thickness	Spatial dependencies of signals	Multiple comparison
SpLoc	✗	√	FWER (in the weak sense)
ST-LME	√	✗	FDR
V-LME	✗	✗	FDR

Finally, we emphasize that we do not obtain a p-value for each cluster under consideration. Since we use the selection algorithm after rejecting the global null hypothesis, the results from the cluster selection algorithm determines the spatial locations that drove the signal and led to the rejection of the global null hypothesis.

2.3. Summary

The proposed method, SpLoc, can be considered as an extension of V-LME in analyzing longitudinal MRI data. V-LME fits a LME (e.g., Eq. (2)) to all vertices and controls FDR after obtaining vertex-wise p-values. A main advantage of V-LME is an explicit modeling of temporal covariance structure by adding subject-specific random effects (Bernal-Rusiel et al., 2013b). SpLoc uses vertex-wise LME as a building block but in a different way. Specifically, we use neighbor information (i.e., Ω) in the hypothesis testing framework to improve statistical power while controlling FWER in a weak sense. Please see Groppe et al. (2011) that compared different methods for controlling for false positives in massive univariate analysis and cluster-wise inference. As a special case, when $\Omega = \{1\}$, SpLoc reduces to the vertex-wise inference and the conceptual difference between SpLoc with $\{1\}$ and V-LME is the nature of adjusting for mutlple comparisons (FWER for SpLoc and FDR for V-LME). Comparing SpLoc to ST-LME, we note that both use spatial information for inferences but their usages are different. ST-LME uses spatial covariance to model the cortical thickness data, hence reducing the standard error of the massive-univariate test statistic. On the other hand, SpLoc specifies a nearest neighbor set for every vertex from the triangulated surface and benefits from adding nearby score test statistics, where 'nearby' is determined data-adaptively through the hypothesis testing framework. The differences of the models considered in this article are summarized in Table 1.

3. Data analysis

3.1. Simulation studies

We conducted simulation studies to evaluate the performance of the proposed method. We used Eq. (1) to generate data. We used the fsaverage4 template to obtain vertices on a triangluated surface of the brain, resulting in approximately 2.5k vertices in each hemisphere. We then computed the pairwise geodesic distance between two vertices and extracted the nearest neighbor information for fitting SpLoc. In each simulation, we fixed the sample size $n = 50$ and let $q = 3$ with $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i1} \cdot x_{i2})^T$, and x_{i1} and x_{i2} were generated using $\mathcal{N}(0, 1)$ and $\text{Binomial}(1, 0.5)$, respectively. A group indicator was randomly drawn for each subject. The number of visits, m_i, was generated using the discrete uniform distribution between 3 and 4 to make sure $m_i > 1$ and we set $\mathbf{t}_i = (0, 0.5, 1, \ldots, 0.5 \cdot (m_i - 1))^T$. The mean parameters $(\alpha_{0k}, \boldsymbol{\alpha}_{1k}, \beta_{0k}, \beta_{1k})$ were fixed as $(1, 1, -1, 0.5, 0.5, 1)$. We used two different setups based on the specifications of the covariance structure, denoted **Simulation 1** and **Simulation 2**:

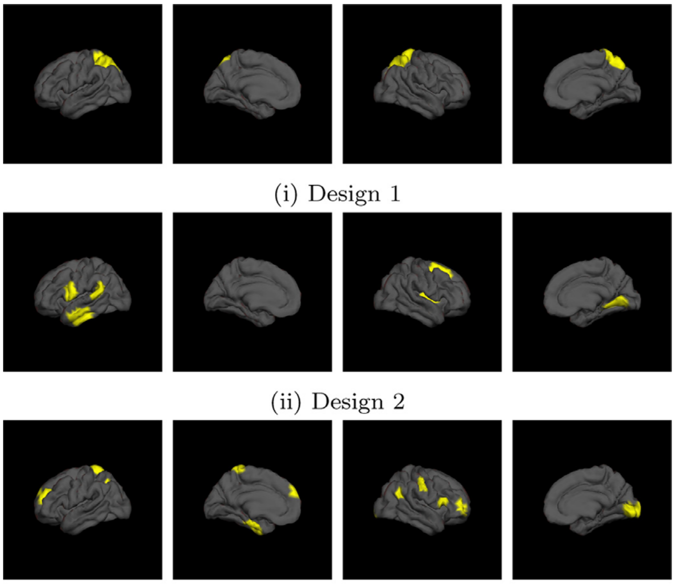
- **Simulation 1**: $\tau_k^2 = 0.5$ and $\boldsymbol{\Phi}_k = \begin{bmatrix} 3 & 0.5 \\ 0.5 & 0.2 \end{bmatrix}$ and there is no spatial correlation. This is a setup where V-LME is statistically more efficient than ST-LME.
- **Simulation 2**: In addition to τ_k^2 and $\boldsymbol{\Phi}_k$ in Simulation 1, we consider spatial correlation structure following Bernal-Rusiel et al. (2013b). This is a setup where ST-LME is statistically more efficient than V-LME. After obtaining the Euclidean distance matrix, we used the exponential spatial correlation structure with spatial correlation parameter 0.1 to generate data. This value was chosen so that each vertex has approximately seven neighboring vertices with a spatial correlation greater than 0.5 and 27 vertices with a spatial correlation greater than 0.3.

We compared SpLoc to V-LME in Simulation 1 and to ST-LME in Simulation 2, based on the efficiency of the competing model due to the simulation setup. For V-LME, we fitted a LME for each vertex and used the Satterthwaite-based approximation to compute vertex-wise p-values (Kenward and Roger, 1997). We then applied two-stage FDR adjustments for multiple testing to obtain q-values (FDR-corrected p-value) (Benjamini et al., 2006). Note that it is appropriate to use the two-stage approach because we used independent (non-spatial) noise in the simulation settings. We used R packages `lme4` (`lmer()` function) for fitting LME models, `lmerTest` (`anova()` function) for the Satterhwaite's approximation, and `mutoss` (`two.stage()` function) for adjusting FDR with the two-stage process (Bates et al., 2014; Benjamini et al., 2006). For ST-LME, we used the MATLAB functions embedded in Freesurfer. For SpLoc, because the number of vertices in fasverage4 is approximately a quarter of the number of vertices in fsaverage5 used in our data analysis, we kept the maximum cluster size to be 250 (a quarter of 1000) so that $\Omega = \{1, 5, 10, \ldots, 250\}$.

We considered three designs for illustrations. In each design, all signal clusters are drawn from the candidate clusters specified by Ω. The visualizations of the designs are provided in Fig. 2. Specifically, in Design 1, we chose a cluster with size 150 in each hemisphere. In Design 2, we constructed three different clusters with size 50 in each hemisphere. In Design 3, we constructed 5 different clusters with sizes 10,20,30,40,50 in each hemisphere. The total number of signal vertices in the brain is the same across the designs (300). For each signal location, we considered different levels of the effect size $\gamma > 0$.

3.2. Simulation results

For each simulation design and setup, we averaged the performance across 1000 simulations. We used two criteria to evaluate performances: power for H_0 and signal detection rate. Power is the number of the simulated datasets we reject the global null hypothesis (i.e., $T_{SpLoc} > t^*$) divided by the total number of simulations (1,000). Signal detection rate for each simulated data is the number of true positives divided by the number of signal locations (300 in each simulation design). We then averaged detection rates across 1000 simulations. For V-LME/ST-LME, we rejected the global null hypothesis when at least one q-value is less than α and the identified vertices are the ones with q-values less than α. We note that FDR is equivalent to FWER when the global null H_0 is true, in which comparing V-LME/ST-LME to SpLoc is fair (Benjamini and Hochberg, 1995). We used $\alpha = 0.05$ throughout the simulation studies.


(i) Design 1

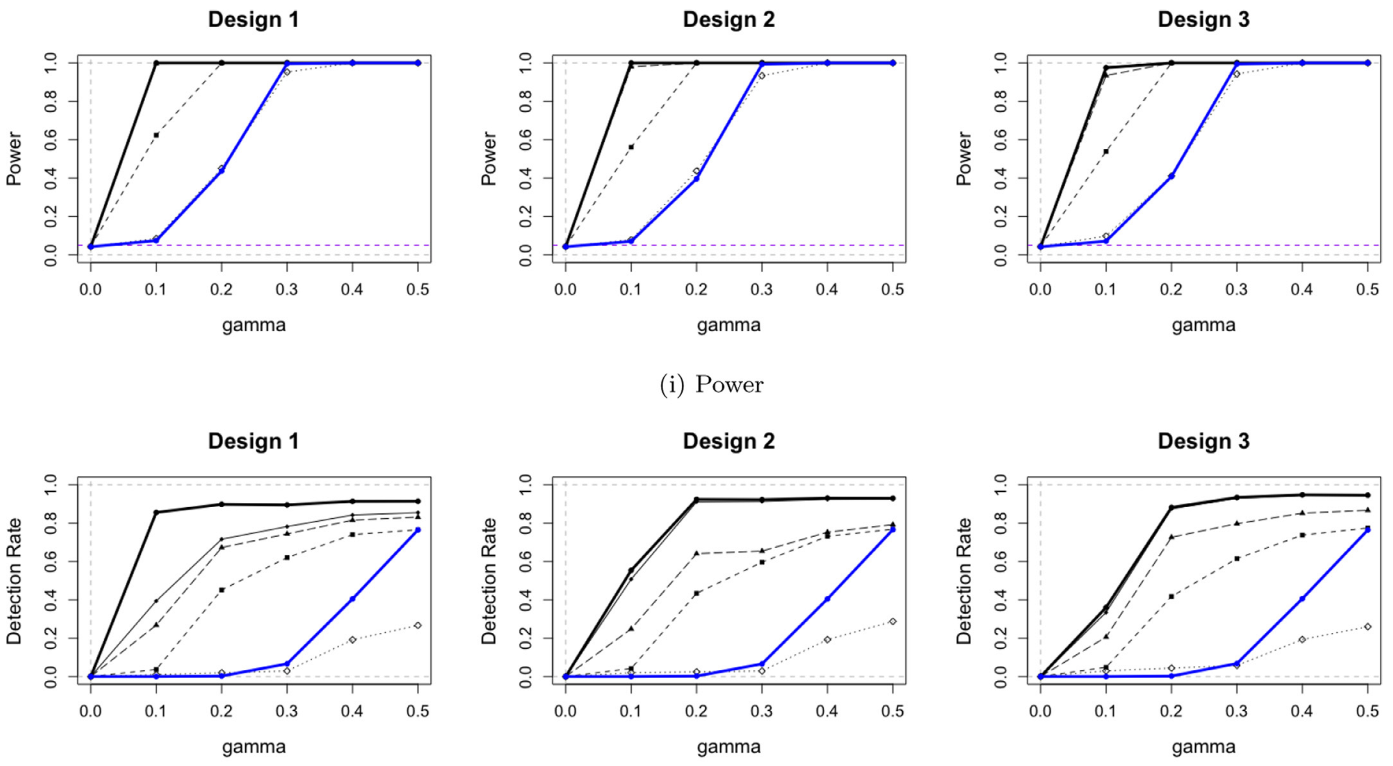
(ii) Design 2

(iii) Design 3

Fig. 2. Illustration of 3 simulation designs. We used fsaverage4 to generate a triangulated surface of the brain. In each hemisphere, Design 1 has a large cluster (150), Design 2 has three clusters of size 50, and Design 3 has 5 clusters with varying sizes $(10, 20, 30, 40, 50)$. The total number of signal vertices is the same across the designs.

The results for Simulation 1 are summarized in Fig. 3. SpLoc (with different choices of Ω) and V-LME all controlled family-wise error properly, with empirical FWER of 0.047 for SpLoc and 0.048 for V-LME. From the perspective of statistical power, Sploc with $\Omega = \{1\}$ nearly had power similar to V-LME, which is not surprising because these methods are using vertex-wise inference but with different methods for adjusting for multiple comparisons. Also, these two are the methods that do not use spatial information, so the power curves were similar in all simulation designs where the equal number of signal vertices was used (300). As the maximum size of the candidate cluster increases, the power of SpLoc increased dramatically but it did not yield noticeable differences in power when the maximum size of the candidate cluster is greater than 50. From the perspective of the signal detection rate, we first see that there are noticeable differences between SpLoc with $\Omega = \{1\}$ and V-LME, and V-LME outperformed SpLoc with $\Omega = \{1\}$ when the effect size gradually increased. This is explained by the difference between FDR and FWER in vertex-wise inferences. However, SpLoc with different Ωs showed superior performance over V-LME. The superior performance is partially due to the power because a candidate cluster, when it correctly captured the true signal cluster, would lead to the increase in power and it would be chosen at the cluster selection step. It is also noticeable from the detection rates that the performance of SpLoc is also affected by the sizes of the true clusters. Provided that we used sufficiently large candidate clusters in Ω, SpLoc had the highest detection rates in Design 1 where all signal vertices formed a large cluster. The detection rates for SpLoc, especially when γ is small, were lowest in Design 3 when there were a few signal clusters with relatively small sizes. Because SpLoc uses the $\max T$ approach, the performance of SpLoc will be driven by the *largest* signal cluster when all clusters are spatially distinct enough. This implies that the smaller signal clusters in Design 3 are relatively not important factors for the power of the SpLoc.

The results for Simulation 2 are summarized in Fig. 4. When spatial correlation was present, SpLoc and ST-LME were conservative in con-



(i) Power

(ii) Detection rate

Fig. 3. Summary of (i) power and (ii) detection rate for **Simulation 1**. The solid black and blue lines denote the results for SpLoc with Ω and V-LME, respectively. The dotted lines are the results of SpLoc using different choices of Ω with respect to the maximum cluster size (◇ = 1, ■ = 10, ▲ = 30, ◆ = 50). The purple dotted line in the power curve is the FWER (0.05).

trolling false positives since they had an empirical FWER of 0.035 and 0.01, respectively. When the null hypothesis was not true, there was no noticeable difference in statistical power between SpLoc and ST-LME. However, we observed that the detection rate of SpLoc is higher than ST-LME. This result suggests that SpLoc performs well even when a moderate degree of spatial correlation is present.

To quantify the error rate under the partial null hypothesis, we averaged the number of false positive vertices divided by the total number of truly null vertices (the total number of vertices minus the number of signal vertices) across 1000 simulated data, for each setup, design and $\gamma > 0$. These error rates were all below a nominal rate 0.05, ranged between 0.033 and 0.038 in Simulation 1 and between 0.001 and 0.026 in Simulation 2. For illustration, we set $\gamma = 0.5$ from Simulation 1 and, for each vertex, we averaged the proportion that that vertex was selected. As shown in Fig. 5, most of the false positives, whenever they exist, are located near the signal vertices. This differs from the result of V-LME/ST-LME based on FDR control, in which false positives varied across vertices on the cortical surface that are truly null.

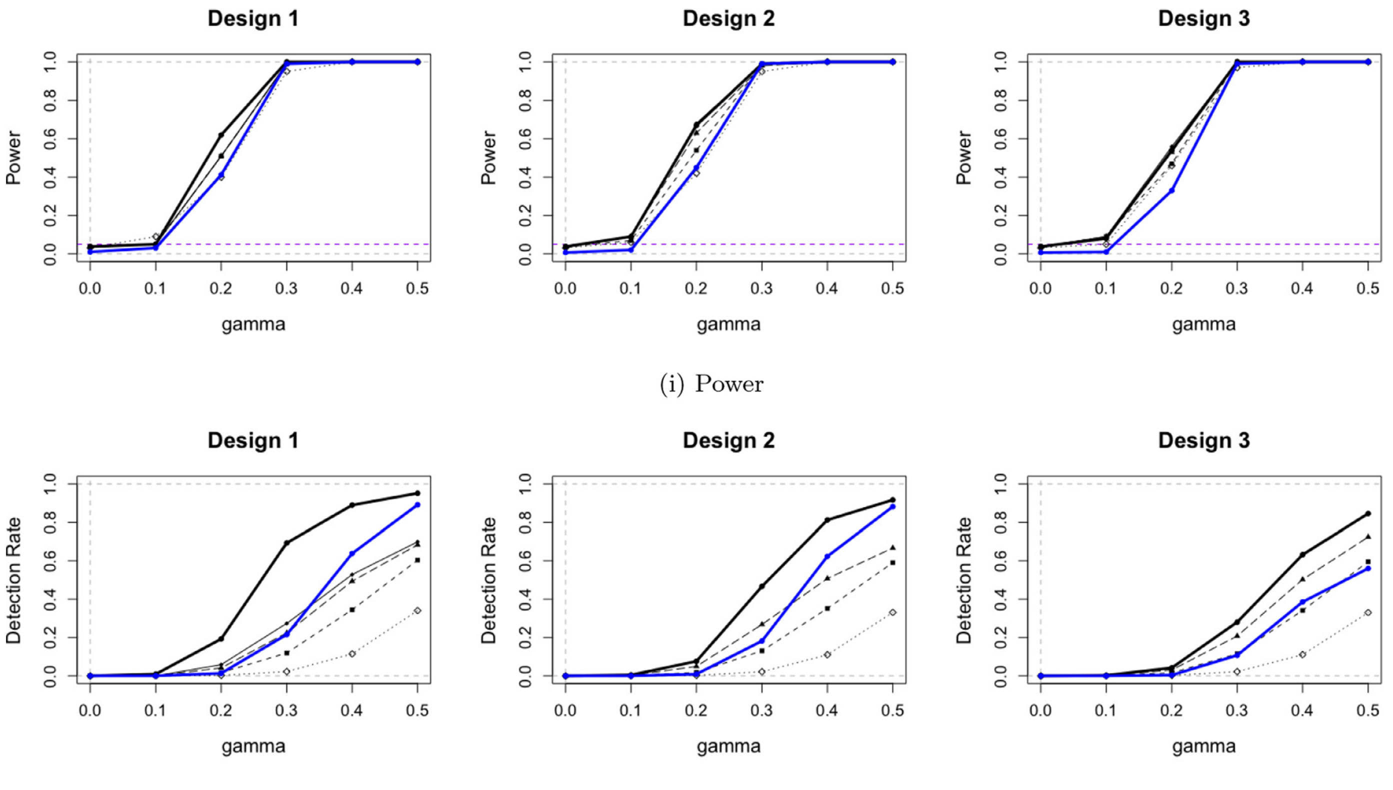
3.3. ADNI Data analysis

Data description

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

We used structural T1-weighted MRI scans obtained from the ADNI-1. We obtained MRI data of 317 subjects by including 127 CN and 190 subjects diagnosed with AD with at least two MRI scans, where the median and maximum number of scans was 4 and 6 respectively. The cortical thickness data were preprocessed using FreeSurfer, a standard tool for MRI processing and extracting cortical thickness. In our work, we employed a minimal spatial smoothing (FWHM = 2mm) so that we do not excessively contaminate the spatial structure in the data with the spatial smoothing due to preprocessing. We did not consider unsmoothed data in our analysis because current practices use smoothing to reduce inter-subject anatomical variations, and a previous work showed that analysis using unsmoothed data has low statistical power (Kang et al., 2015). All MR images were registered to the fsaverage5 template, a downsampled version of the fsaverage template, resulting in 10,242 vertices in each hemisphere. Among those, vertices in the corpus callosum, defined by the Desikan-Killiany Atlas (Desikan et al., 2006), are discarded as it is non-cortical. We then used the geodesic distance to compute pairwise distance between two vertices and obtain set of nearest neighbors (Kirsanov, 2008).

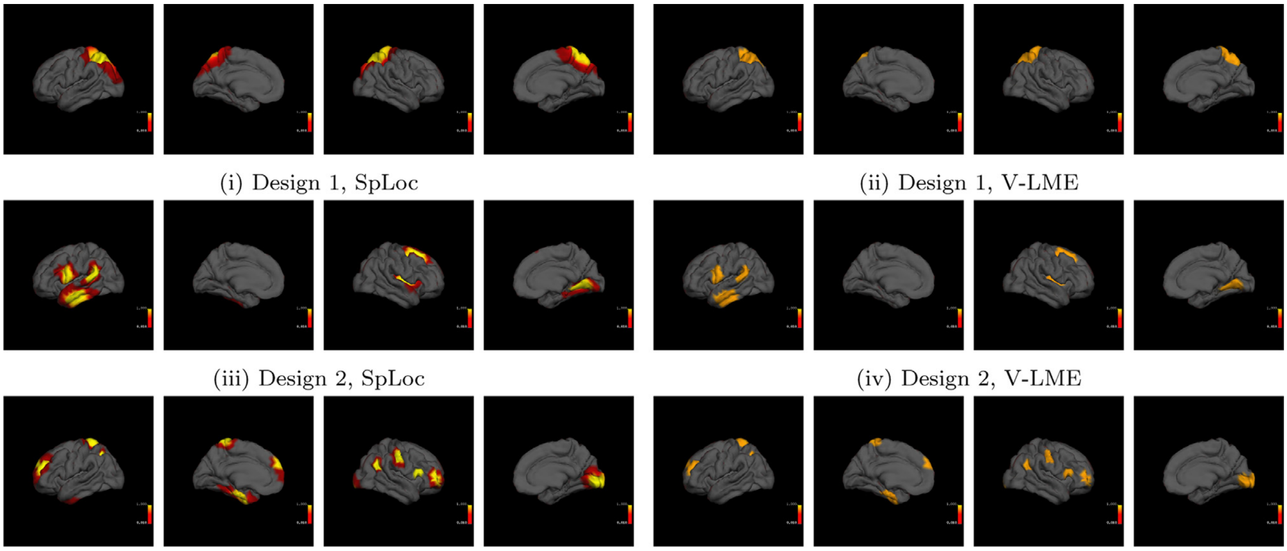
The baseline covariate information we used in our analysis was obtained from the R package ADNIMERGE. We considered baseline age, gender, and number of years of education, time (from the baseline), APOE genotype status (one if e4 carrier and 0 else), and the total brain volume as covariates. SpLoc as well as V-LME and ST-LME are based on the LME framework, and we used random intercept and random slope of time as random effects. Both V-LME and ST-LME are available in FreeSurfer, and we used the recommended options to implement both models. Specifically, we used an exponential correlation structure with Euclidean distance to define the pairwise spatial correlation between two vertices and used $k = 2$ (defined by Bernal-Rusiel et al. (2013b)) to define region parcellations for ST-LME.



(i) Power

(ii) Detection rate

Fig. 4. Summary of (i) power and (ii) detection rate for **Simulation 2**. The solid black and blue lines denote the results for SpLoc with Ω and ST-LME, respectively. The dotted lines are the results of SpLoc using different choices of Ω with respect to the maximum cluster size (◇ = 1, ■ = 10, ▲ = 30, ◆ = 50). The purple dotted line in the power curve is the FWER (0.05).



(i) Design 1, SpLoc

(ii) Design 1, V-LME

(iii) Design 2, SpLoc

(iv) Design 2, V-LME

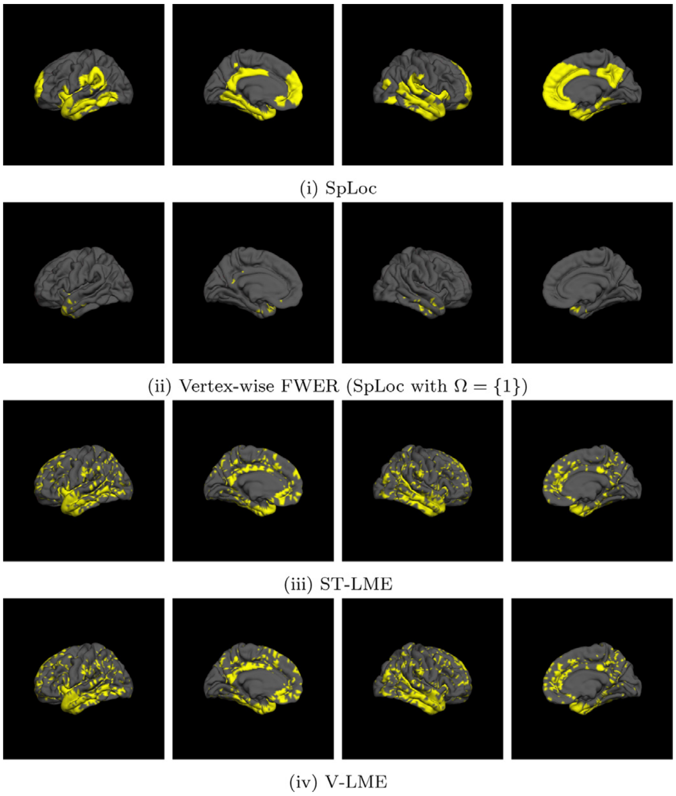
(v) Design 3, SpLoc

(vi) Design 3, V-LME

Fig. 5. Visualizations of the **Simulation 1** when $\gamma = 0.5$. For each vertex, we computed the proportion selected by each method across 1000 simulated data.

Results

Fig. 6 illustrates the vertices detected by SpLoc (with 10,000 permutations), vertex-wise inference controlling FWER (achieved by using SpLoc with $\Omega = \{1\}$), ST-LME and V-LME. SpLoc identified 20 and 25 localized clusters in the left and right hemispheres, with the maximum cluster size of 500 (left) and 800 (right) and the minimum cluster size of 5 (left) and 1 (right). Overall, SpLoc declared statistical significance in 27.2% and 32.4% of the total vertices in the left and right hemispheres, respectively, in the fsaverage5 template. The areas of significance detected by SpLoc are in general agreement with exising work in neuroimaging literature, especially along the inferior and middle temporal regions and the temporal pole in both hemispheres (Bernal-Rusiel et al., 2013b; Dickerson et al., 2009; Eskildsen et al., 2013). Using the Desikan-Killiany Atlas (Desikan et al., 2006), the areas detected by SpLoc cov-



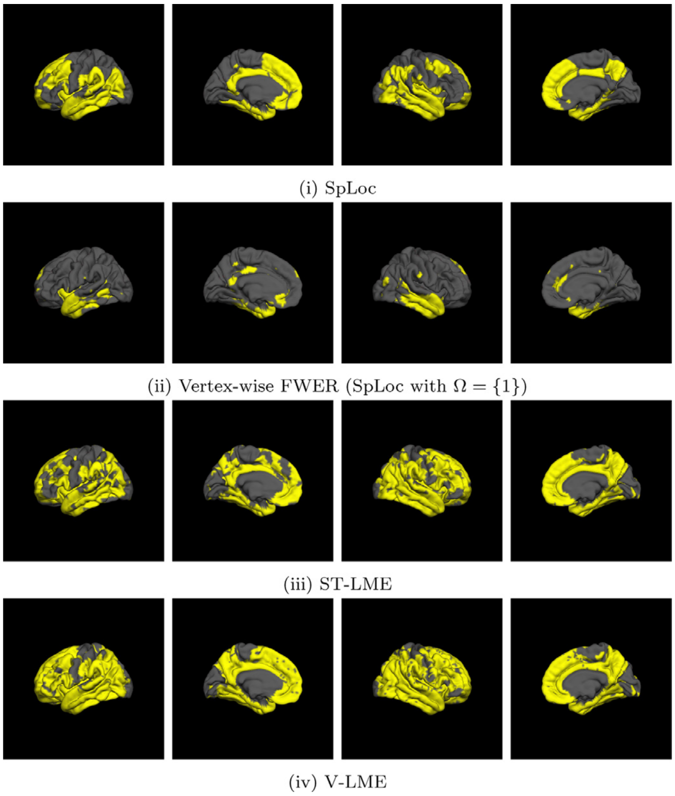
(i) SpLoc

(ii) Vertex-wise FWER (SpLoc with $\Omega = \{1\}$)

(iii) ST-LME

(iv) V-LME

Fig. 6. Illustration of selected vertices for each model using a low level of smoothing (FWHM= 2mm). Each column denotes lateral view (left), medial view (left), lateral view (right), and medial view (right), respectively.



(i) SpLoc

(ii) Vertex-wise FWER (SpLoc with $\Omega = \{1\}$)

(iii) ST-LME

(iv) V-LME

Fig. 7. Illustration of selected vertices for each model using a wide range of smoothing (FWHM= 10mm). Each column denotes lateral view (left), medial view (left), lateral view (right), and medial view (right), respectively.

ered more than 80% of the areas of the entorhinal, frontal pole, parahippocampal and temporal pole cortices. In addition, the left hemisphere covered the inferior temporal, isthmus cingulate, and posterior cingulate cortices and the right hemisphere covered the caudal and rostral anterior cingulate, insula, and medial orbitofrontal cortices.

Comparing the methods, SpLoc overcame the conservative nature of vertex-wise inference with FWER control. It is also noticeable that ST-LME and V-LME do not have meaningful differences in identified vertices. This is not surprising because the median number of vertices in each partition was 2, which is insufficient to benefit from adding a spatial covariance. Both ST-LME and V-LME identified parts of the brain with significant cortical thinning rate differences, but the areas of significance were substantially smaller than the areas detected by SpLoc.

We conducted sensitivity analysis to see if the areas of significance detected by SpLoc vary by changing the numbers of maximum cluster size. We first increased the maximum cluster size from 1000 to $1100, 1200, \dots, 2000$ and the vertices detected by the regions remained the same. Similar to the simulation results, the number of detected vertices gradually increased when we gradually increased the maximum cluster size from 1, and the number is maximized when the maximum cluster sizes were 500, 600, and 700. However, there was no difference in the left hemisphere and minor differences in the right hemisphere and 96.3% of the vertices detected by the original analysis were replicated. This result suggests that the result of SpLoc is not significantly affected by the choice of maximum cluster size, provided that it is big enough to capture large signal clusters of the brain.

We also applied SpLoc and competing methods to the same dataset but with a different smoothing level (FWHM = 10mm) during preprocessing to evaluate the impact of spatial correlation on inference. The results using the same methods are shown in Fig. 7. Putting aside the issue of spatial specificity caused by a wide-range of smoothing, ST-LME and V-LME detected more vertices than SpLoc, and vertex-wise inference controlling FWER still suffered from detecting signal regions identified in previous neuroimaging literature. The proportion of vertices detected by SpLoc increased to 42.2% (left) and 46.6% (right). SpLoc detected 188 (left) and 215 (right) clusters but 161 and 188 of them consisted of a single vertex only. The result suggests that ST-LME and V-LME with FDR adjustments can be more useful when spatial correlation is present. This result is not surprising, because SpLoc uses vertex-wise models for conducting spatial-extent inferences, which might lose statistical efficiency.

Altogether, from the data analysis we saw that SpLoc is a useful alternative to the massive univariate analysis when the signal clusters are expected to be big and spatial correlation is relatively low. SpLoc is particularly useful in cortical thickness data where a wide range of smoothing is needed to decrease noise level of each vertex at the expense of the loss of spatial specificity. Finally, SpLoc successfully recovered most of the regions previously identified in the neuroimaging literature, even though the smoothing level during preprocessing is minimal.

4. Discussion

We proposed a novel inference procedure for analyzing longitudinal neuroimaging data, called SpLoc, that detects spatial clusters of vertices where the rate of cortical atrophy differs between two cohorts. Our work is summarized in three steps: (i) fitting univariate (vertex-wise) LME and data-adaptively combining combining test statistics across neighboring vertices, (ii) proposing a computationally efficient permutation-based inference to control FWER and estimate the covariance of the univariate test statistics, and (iii) selecting multiple spatial clusters by pruning off multiple candidate clusters based on the ordering of the cluster-wise test statistics. The proposed method is based on the LME framework that is widely used for analyzing longitudinal neuroimaging data, and it is a simple yet powerful approach that uses the vertex-wise LME results for detecting spatial clusters. We show using both simulation studies and an analysis of the ADNI data that the proposed method outperforms existing methods.

Existing works for analyzing cortical data generally smoothed the cortical thickness data using a Gaussian kernel with FWHM at least 8mm, which improved their statistical power at the expense of spatial specificity. In our data analysis, we used a minimal level of smoothing (FWHM= 2mm) during preprocessing so that our inference was driven more by the spatial extent of the signal in the data and not by the smoothing due to preprocessing. Our data analysis showed that SpLoc is a powerful approach that maintains FWER control for multiple comparisons, and this is accomplished by adaptively combining neighbor information without relying on strict distributional assumptions. We believe, however, that an optimal level of smoothing during preprocessing needs to be determined since this can result in a tradeoff between sensitivity and specificity. Alternatively, one could proceed with analyzing the data using data smoothed at different levels, as we had done in this work.

We describe some limitations of SpLoc. As shown in Fig. 1 and in our simulation study, the choice of the neighbor set can affect the sensitivity and specificity. While we showed in our simulations that SpLoc maintained favorable power and FWER control in the weak sense relative to V-LME, Ω needs to be picked carefully. Because the family-wise error is controlled in the weak sense in cluster-wise inference, statistically significant regions detected by SpLoc may suffer from the loss of specificity. A follow-up study on specificity that compares SpLoc to other cluster-wise inference methods would be helpful.

Our proposed method has room for improvement in several ways. First, similar to how ST-LME outperforms V-LME, adopting the region/brain-wise parametric spatial covariance matrix in to the framework may improve statistical power (Bernal-Rusiel et al., 2013b; Bowman, 2007; Kang et al., 2012). Also, accounting for subject-level heterogeneity of spatial covariance structures in robust estimation would be worth considering in MRI literature (Vandekar et al., 2019). However, spatial modeling in a generalized linear model (GLM) framework can be computationally intensive especially when the number of vertices is large. A form of dimension reduction or approximation methods would be necessary to maintain reasonable computational cost in adding a spatial structure to the model. Second, the construction of nearest neighbors in this article is based on geodesic distance only, and it is limited in approximating the shape of true signal clusters. Third, even though it is not expected to change the overall message of this article, it is worth considering other factors that would yield more accurate results in analyzing longitudinal cortical thickness data. As an example, the ADNI data used in our analysis was collected from different scanners and sites, and we believe the harmonization of cortical thickness across scanners and sites would remove unwanted sources of variation in the LME framework (Beer et al., 2020; Fortin et al., 2018). Also, there is evidence of cortical thinning of select regions of the brain to accelerate over time, and including a nonlinear time effect to capture this phenomenon may improve model fit to the data (Bilgel et al., 2016; Fjell et al., 2014; McDonald et al., 2009).

A key advantage of SpLoc is that it is not limited to the specific LME model specified in this article. Even though we were primarily interested in finding regions with different cortical decay rates in longitudinal MRI, it is possible to use SpLoc to test for main effects for clinical status by setting the null hypothesis for the appropriate parameters in the linear model and then using an appropriate permutation strategy. Also, it can be extended to test hypothesis in different linear model frameworks. An interesting future work is to extend SpLoc to other neuroimaging data types. We are currently working on applying SpLoc to group-level activation in task-based fMRI where a number of cluster-wise inference methods have been proposed.

R code for implementing SpLoc, as well as documented examples used in simulation studies, is publicly available as a form of R package at https://github.com/junjypark/SpLoc.

Declaration of Interest

None.

Credit authorship contribution statement

Jun Young Park: Conceptualization, Methodology, Data curation, Software, Formal analysis, Validation, Visualization, Writing - original draft. **Mark Fiecas:** Conceptualization, Methodology, Data curation, Visualization, Writing - review & editing.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The authors appreciate the editor and anonymous reviewers for their helpful and constructive comments. Also, the authors would like to thank Ning Dai (Ph.D. Statistics, University of Minnesota) and Jin Jin (Ph.D. Biostatistics, University of Minnesota) for their helpful comments on data preprocessing. The computing resources were supported by the Minnesota Supercomputing Institute (MSI). Jun Young Park was supported by the MnDRIVE Informatics PhD Graduate Assistantship Program at the University of Minnesota.

Appendix A. Illustration of the cluster selection algorithm

Consider a 4 × 4 grid of spatial locations, and suppose that the true, unknown signal clusters are located as shown in Fig. 8. Our cluster selection algorithm proceeds as follows:

Step 1. Compute the test statistic for candidate cluster (i.e., $T_k^{(r)}$) we defined, using the fitted model under the null hypothesis H_0. In our example, we defined candidate clusters as square blocks of different sizes as shown in Fig. 9.

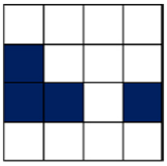


Fig. 8. True signal locations.

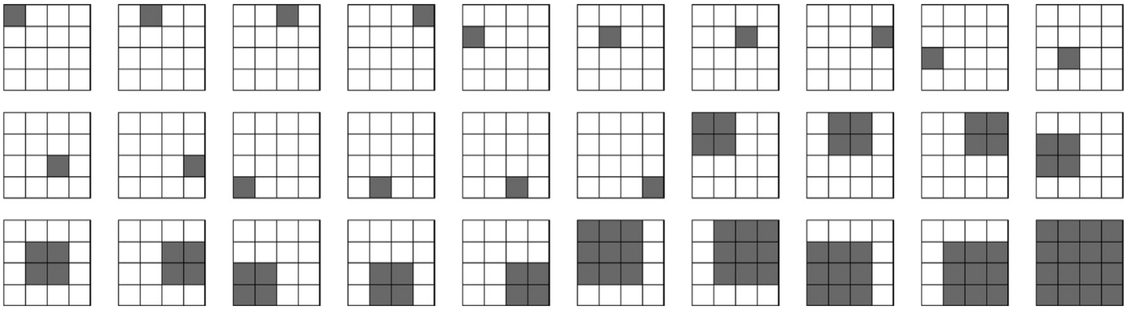


Fig. 9. 30 candidate clusters defined by square blocks of different sizes.

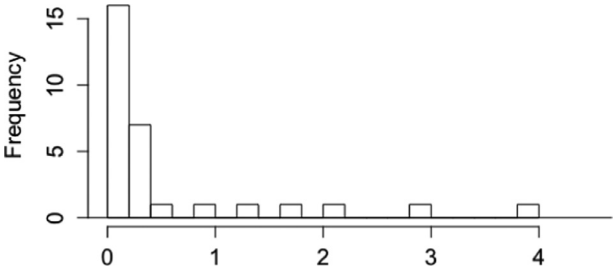


Fig. 10. Histogram of test statistics for every candidate cluster.

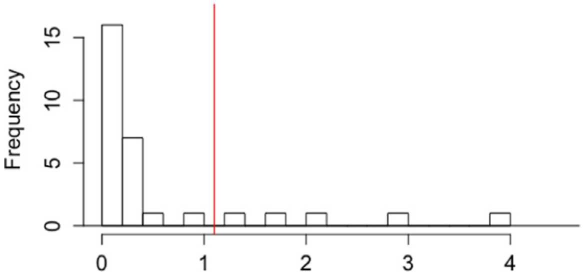


Fig. 11. Histogram of test statistics for every candidate cluster, with a threshold that controls FWER.

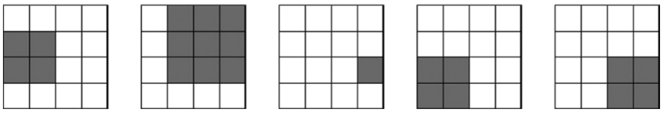


Fig. 12. Five candidate clusters whose test statistics are greater than a threshold that controls FWER.

The histogram of the test statistics can be obtained as shown in Fig. 10.

Step 2. Using permutation, compute the threshold t^* that controls FWER at the rate of α (e.g., 0.05). The threshold in this example is shown in the red line in Fig. 11.

Step 3. If T_{SpLoc} is less than t^*, conclude that there is no signal location. Otherwise (as shown above), ignore all candidates whose test statistics $T_k^{(r)}$ are less than t^* and collect the candidates, as shown in Fig. 12 (ordered by test statistic, from the largest).

Step 4. Identify the candidate with the largest value as a signal location. Then erase all the other candidates that overlap with the cluster. This step will erase the second (3×3) and the fourth (2×2) candidates.

Step 5. Identify the candidate with the largest value as a signal location, after excluding the pre-defined signal locations and the erased candidates. This step will identify the third (1×1) candidate as a signal.

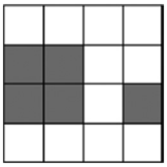


Fig. 13. Identified Signal locations.

Step 6. Repeat Step 5 until there is nothing left. This step will additionally erase the last (2×2) candidate and there's no other candidate left.

Step 7. Combine all identified signal clusters. The identified signal clusters are shown in Fig. 13.

References

Bates, D., Mächler, M., Bolker, B., Walker, S., 2014. Fitting linear mixed-effects models using lme4. arXiv:1406.5823.

Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., Linn, K.A., Initiative, A.D.N., et al., 2020. Longitudinal ComBat: a method for harmonizing longitudinal multi-scanner imaging data. Neuroimage 220, 117129.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57 (1), 289–300.

Benjamini, Y., Krieger, A.M., Yekutieli, D., 2006. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93 (3), 491–507.

Bernal-Rusiel, J.L., Atienza, M., Cantero, J.L., 2010. Determining the optimal level of smoothing in cortical thickness analysis: a hierarchical approach based on sequential statistical thresholding. Neuroimage 52 (1), 158–171.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., Initiative, A.D.N., et al., 2013. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. Neuroimage 66, 249–260.

Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., Sabuncu, M.R., Initiative, A.D.N., et al., 2013. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. Neuroimage 81, 358–370.

Bilgel, M., Prince, J.L., Wong, D.F., Resnick, S.M., Jedynak, B.M., 2016. A multivariate nonlinear mixed effects model for longitudinal image analysis: application to amyloid imaging. Neuroimage 134, 658–670.

Bowman, F.D., 2007. Spatiotemporal models for region of interest analyses of functional neuroimaging data. J. Am. Stat. Assoc. 102 (478), 442–453.

Braun, T.M., Feng, Z., 2001. Optimal permutation tests for the analysis of group randomized trials. J. Am. Stat. Assoc. 96 (456), 1424–1432.

Buse, A., 1982. The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note. Am. Stat. 36 (3a), 153–157.

Chan, D., Janssen, J.C., Whitwell, J.L., Watt, H.C., Jenkins, R., Frost, C., Rossor, M.N., Fox, N.C., 2003. Change in rates of cerebral atrophy over time in early-onset Alzheimer's disease: longitudinal MRI study. Lancet 362 (9390), 1121–1122.

Coalson, T.S., Van Essen, D.C., Glasser, M.F., 2018. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. Proc. Natl. Acad. Sci. 115 (27), E6356–E6365.

Cox, D.R., Hinkley, D.V., 1979. Theoretical Statistics. CRC Press.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31 (3), 968–980.

Dickerson, B.C., Bakkour, A., Salat, D.H., Feczko, E., Pacheco, J., Greve, D.N., Grodstein, F., Wright, C.I., Blacker, D., Rosas, H.D., et al., 2009. The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb. Cortex 19 (3), 497–510.

Draper, N.R., Stoneman, D.M., 1966. Testing for the inclusion of variables in linear regression by a randomisation technique. Technometrics 8 (4), 695–699.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., Initiative, A.D.N., et al., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. Neuroimage 65, 511–521.

Fischl, B., 2012. Freesurfer. Neuroimage 62 (2), 774–781.

Fjell, A.M., Westlye, L.T., Grydeland, H., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Dale, A.M., Walhovd, K.B., Initiative, A.D.N., 2014. Accelerating cortical thinning: unique to dementia or universal in aging? Cereb. Cortex 24 (4), 919–934.

Fortin, J.-P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.

Ganjgahi, H., Winkler, A.M., Glahn, D.C., Blangero, J., Kochunov, P., Nichols, T.E., 2015. Fast and powerful heritability inference for family-based neuroimaging studies. Neuroimage 115, 256–268.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15 (4), 870–878.

Gordon, B.A., Blazey, T.M., Su, Y., Hari-Raj, A., Dincer, A., Flores, S., Christensen, J., McDade, E., Wang, G., Xiong, C., et al., 2018. Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. Lancet Neurol. 17 (3), 241–250.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011. Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. Psychophysiology 48 (12), 1711–1725.

Guillaume, B., Hua, X., Thompson, P.M., Waldorp, L., Nichols, T.E., Initiative, A.D.N., et al., 2014. Fast and accurate modelling of longitudinal and repeated measures neuroimaging data. Neuroimage 94, 287–302.

Hart, B., Cribben, I., Fiecas, M., Initiative, A.D.N., et al., 2018. A longitudinal model for functional connectivity networks using resting-state fMRI. Neuroimage 178, 687–701.

Jeng, X.J., Cai, T.T., Li, H., 2010. Optimal sparse segment identification with application in copy number variation analysis. J. Am. Stat. Assoc. 105 (491), 1156–1166.

Kang, H., Blume, J., Ombao, H., Badre, D., 2015. Simultaneous control of error rates in fmri data analysis. Neuroimage 123, 102–113.

Kang, H., Ombao, H., Linkletter, C., Long, N., Badre, D., 2012. Spatio-spectral mixed-effects model for functional magnetic resonance imaging data. J. Am. Stat. Assoc. 107 (498), 568–577.

Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 983–997.

Kim, J., Wozniak, J.R., Mueller, B.A., Shen, X., Pan, W., 2014. Comparison of statistical tests for group differences in brain functional networks. Neuroimage 101, 681–694.

Kirsanov, D., 2008. geodesic: Multiple source/target exact geodesic (shortest path) algorithm for triangular mesh (triangulated 2D surface in 3D).

Landin-Romero, R., Kumfor, F., Leyton, C.E., Irish, M., Hodges, J.R., Piguet, O., 2017. Disease-specific patterns of cortical and subcortical degeneration in a longitudinal study of Alzheimer's disease and behavioural-variant frontotemporal dementia. Neuroimage 151, 72–80.

Lee, S., Abecasis, G.R., Boehnke, M., Lin, X., 2014. Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 95 (1), 5–23.

Lerch, J.P., Evans, A.C., 2005. Cortical thickness analysis examined through power analysis and a population simulation. Neuroimage 24 (1), 163–173.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.

McDonald, C., McEvoy, L., Gharapetian, L., Fennema-Notestine, C., Hagler, D., Holland, D., Koyama, A., Brewer, J., Dale, A., et al., 2009. Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. Neurology 73 (6), 457–465.

Staffaroni, A.M., Brown, J.A., Casaletto, K.B., Elahi, F.M., Deng, J., Neuhaus, J., Cobigo, Y., Mumford, P.S., Walters, S., Saloner, R., et al., 2018. The longitudinal trajectory of default mode network connectivity in healthy older adults varies as a function of age and is associated with changes in episodic memory and processing speed. J. Neurosci. 38 (11), 2809–2817.

Vandekar, S.N., Satterthwaite, T.D., Xia, C.H., Adebimpe, A., Ruparel, K., Gur, R.C., Gur, R.E., Shinohara, R.T., 2019. Robust spatial extent inference with a semiparametric bootstrap joint inference procedure. Biometrics 75 (4), 1145–1155.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397.

Xu, Z., Shen, X., Pan, W., Initiative, A.D.N., et al., 2014. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. PLoS ONE 9 (8), e102312.